Journal of Inequalities and Applications
a SpringerOpen Journal

RESEARCH Open Access

Wavelet estimation of copula function based on censored data

Bahareh Ghanbari[1], Masoud Yarmohammadi[1*], Narges Hosseinioun[2] and Esmaeil Shirazi[3]

*Correspondence:
Yarmohammadi.mas@gmail.com
[1] Department of Statistics, Payame Noor University, Tehran, Iran
Full list of author information is available at the end of the article

Abstract

In this paper, we consider wavelet analysis to obtain an estimator of a copula function based on censored data. We show that optimal convergence rates for the mean integrated squared error (MISE) of linear wavelet-based function estimators are exact under right censoring model. Moreover, we derive asymptotic formulae for MISE. Finally, the simulation results and the analysis of real data validate the proposed procedure.

Keywords: Censoring; Copula function; Kaplan–Meier estimator; Nonparametric estimation; Ranks; Wavelet estimators

1 Introduction

Recently, copulas and their applications in statistics have become a rather popular phenomenon. For a long time, statisticians have been interested in the relationship between a multivariate distribution function and its lower-dimensional margins. When it comes to analyzing the dependence between random variables, the copula function becomes a very useful tool. According to Sklar's theorem [22], if H is a bivariate distribution function with margins $F(x)$ and $G(y)$, then there exists a copula C such that $H(x,y) = C(F(x), G(y))$. Statisticians are interested in copulas for two reasons: (1) Inspecting how the measures of dependence are scale-free and (2) constructing families of bivariate distributions. For more information about this flexible tool and dependence modeling, we refer to [16] and [20]. A detailed survey on copula models can be found in [24]. Copula models have many applications in finance and insurance, some of which are considered in [2, 7, 10, 11, 13]. For more studies on copulas in econometrics, we refer to [3] and [21].

In recent years, the analysis of censored data has become very popular. In many sciences such as statistics, engineering, finances, and other areas, censoring is a condition in which observation is partially known, that is, censoring occurs when a value is outside the range of a measuring instrument, particularly in life-testing data in medicine, where observations often survive or fail at the end of the test (see, e.g., [12] for good examples of the censoring method). Let $T_1, T_2, \dots, T_n$ be i.i.d. survival times with probability density function f and common distribution function F. Typically, instead of observing the variables of interest, we are able to observe the i.i.d. censoring times $C_1, C_2, \dots, C_n$. Let G be the common distribution function of the latter. Also, it is assumed that T_i and C_i are independent. Let $Y_i = \min(T_i, C_i)$ for $i = 1, \ldots, n$, and define the indicator function

$$\delta_i = \begin{cases} 1 & \text{if } T_i \le C_i, \\ 0 & \text{if } T_i > C_i. \end{cases}$$

This definition denotes the right-censoring. Several approaches have been proposed to deal with the estimation of copulas under uncensored or censored data. See [9] for nonparametric estimation and also [23]. In [24] consistent estimators are considered under some restrictions on the dependence structure of the censored data. For right-censored data, less work has been done on nonparametric copula estimators. Recently, in [15] a new class of nonparametric estimators of copula function for bivariate censoring is described. The aim of this paper is to estimate the copula function via wavelets based on censoring.

The theory of wavelets and their applications in statistics and other sciences have become an important technique. We can find several applications of wavelet estimators for copula functions in different contexts. In [14] smoothed empirical copula estimators considered using ranks and wavelets. See [1] for more information about the procedure to estimate copula functions using wavelets, where the problem of multivariate copula density estimation by wavelet methods has been described. In [19] the wavelet approaches were used, basically following the same route as in [14]. In [5] multiwavelets for estimating copula density were used, and in [6] the same work was prepared based on a Legendre multiwavelets. There are many different papers that considered wavelets, and we proposed some of the most important.

Currently, an extension of wavelets and copulas is increasingly popular as an alternative to many methods such as those briefly discussed. In this paper, we propose a linear wavelet-based estimation for copulas with right censored data in the observation T_1 or T_2. Let $(T_{11}, T_{21}), \ldots, (T_{1n}, T_{2n})$ be independent and have a joint distribution function (d.f.) $F(t_1, t_2)$. As usual, assume that T_1 and C_1 are independent, where C_1 is the censoring variable associated with the variable T_1. Rather than observing the variables of interest $(T_{11}, T_{21}), \ldots, (T_{1n}, T_{2n})$, in the randomly right censor model, (Y_i, δ_i, T_{2i}) is observed. We apply the method of [17], which provides a MISE expansion similar to a density function over $(-\infty, T]$ for any fixed $T < \tau_{H(\cdot,t)}$, where $\tau_{H(\cdot,t)} = \inf\{u : H(u,t) = 1\} \le \infty$ is the least upper bound for the support of $H(\cdot, t)$, the distribution function of (Z_1, t).

The rest of the paper is as follows. In Sect. 2, we define the elements of copula density and wavelet transform and provide the wavelet estimators for the copula based on the censored data. The main results are described in Sect. 3, and the simulation study for the proposed estimator is provided in Sect. 4. The proofs are presented in the Appendix.

2 Preliminary notations

In this section, we provide some necessary concepts about the general framework of this paper. Section 2.1 presents preliminary assumptions and some notations about copula functions. All the major definitions and facts about the wavelets used in the paper are presented in Sect. 2.2. At the end of the section we introduce the estimators for the main result.

2.1 Copula function

Copula had been first established in [22] and improved very fast in many spaces. Here we give a brief definition of a copula function in the two-dimensional case. Any extension of the results to higher dimensions is straightforward. Consider a random vector (T_1, T_2) with cumulative distribution function $F(t_1, t_2) = P(T_1 \leq t_1, T_2 \leq t_2)$ and margin distribution functions F_1 and F_2. The relation between these variables is of interest, so the copula function C can be formed as follows:

$$F(t_1, t_2) = C\bigl(F_1(t_1), F_2(t_2)\bigr).$$

The marginals F_1 and F_2 have uniform distribution on (0,1), and if they are continuous, then C is unique and coincides with the distribution function of the pair $(U, V) = (F_1(T_1), F_2(T_2))$. In practice, F is unknown. The advantage of using copula is that the joint distribution function $(F(t_1, t_2))$ can be constructed by using marginals (F_1 and F_2) when they are from different classes of distributions. Let $(T_{11}, T_{21}), \ldots, (T_{1n}, T_{2n})$ be a random sample from the unknown distribution F. Denote by F_{1n} and F_{2n} the empirical distributions associated with F_1 and F_2. A first step in selecting an appropriate class of copulas consists of plotting the pairs $(\frac{R_i}{n}, \frac{S_i}{n}) = (F_{1n}(T_{1i}), F_{2n}(T_{2i}))$, $i = \{1, \ldots, n\}$.

Here R_i is the rank of T_{1i} among $T_{11}, \ldots, T_{1n}$, and S_i is the rank of T_{2i} among $T_{21}, \ldots, T_{2n}$. The motivation behind this approach is that the pseudo-observations $(R_i/n, S_i/n)$ are close substitutes to the unobservable pairs $(U_i, V_i) = (F_1(T_{1i}), F_2(T_{2i}))$ forming a random sample from C. We denote by $c(u, v)$ the density of $C(u, v)$,

$$c(u, v) = \frac{\partial^2 C(u, v)}{\partial u \partial v}, \quad u, v \in (0, 1).$$

It is obvious that in real analysis one of our variables (or even both of them) may be subject to be censored, and one observes a minimum between it and another (censoring) random variable, denoted by C_j, so $Y_j = \min(T_j, C_j)$ and $\delta_j = I_{T_j \leq C_j}$ for $j = 1, 2$. The i.i.d. replication vector $(Y_{1i}, Y_{2i}, \delta_{1i}, \delta_{2i})_{1 \leq i \leq n}$ denotes the random sample of $(Y_1, Y_2, \delta_1, \delta_2)$. Clearly, many different estimators can be considered for a distribution function based on censoring. The one used in this paper has the form

$$\hat{F}_n(t_1, t_2) = \frac{1}{n} \sum_{i=1}^{n} W_{in}^{(\cdot)} I(Y_{1i} \leq t_1, T_{2i} \leq t_2),$$

where W_{in} are random weights designed to compensate asymptotically the bias caused by censoring and can be assumed in three different forms; see [15]. The weight considered in the present paper is

$$W_{in} = \frac{\delta_{1i}}{1 - \hat{G}(Y_{1i}^{-})}.$$

In this form only T_1 is assumed to be censored, and then $Y_2 = T_2$, $\delta_2 = 1$ a.s., and C_1 is independent from T_1. Also $\hat{G}$, the Kaplan–Meier estimator of the censoring variable, is defined as $\hat{G}(t) = 1 - \prod_{i: Y_{1i} \leq t} (1 - \frac{1}{\sum_{j=1}^{n} I(Y_{1j} \geq Y_{1i})})^{1 - \delta_{1i}}$. Introducing $G(t) = P(C_1 \leq t)$, the weight W_{in} can be seen as an approximation of $W_i = \frac{\delta_{1i}}{1 - G(Y_{1i}^{-})}$. The other weights are discussed in [15]. They took $W_{in} = \delta_{1i} \hat{g}(Y_{1i})$, where $\hat{g}$ is a consistent estimator of limit function g, estimated from the data, where g satisfies the condition $E[\delta_1 g(Y_1) \phi(Y_1, T_2)] = E[\phi(T_1, T_2)]$ for all $\phi \in L^1$.

2.2 Wavelets

Wavelets and their applications are still an important subject in statistics. The term wavelet is used to refer to a set of orthonormal basis functions generated by dilation and translation of a compactly supported scaling function (father wavelet) ϕ and a mother wavelet ψ associated with an r-regular ($r > 0$) multiresolution analysis of $L_2(\mathbb{R})$, the space of square-integrable functions on the line. Define $\phi_{j,k}(x) = 2^{j/2}\phi(2^j x - k)$ and $\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$ for $j \in \mathbb{N}$ and $k = (k_1, k_2) \in \mathbb{Z}^2$. It is assumed that $j \geq j_o$ for some coarse scale $j_o \in \mathbb{N}$, which we take as l. We suppose that ϕ and ψ are bounded and compactly supported. For more on wavelets, see [8] and [18]. The wavelet expansion for $f(x,y)$ can be written as

$$f(x,y) = f_{j_0}(x,y) + D_{j_0}f(x,y), \quad x,y \in R, \tag{1}$$

where $f_{j_0}(x,y) = \sum_{k\in\mathbb{Z}^2} \alpha_{j_0k}\phi_{j_0k}(x,y)$ is a trend of an approximation, and

$$D_{j_0}f(x,y) = \sum_{j=j_0}^{\infty}\left(\sum_{k\in\mathbb{Z}^2}\beta_{j_0k}^{(1)}\psi_{j_0k}^{(1)}(x,y) + \sum_{k\in\mathbb{Z}^2}\beta_{j_0k}^{(2)}\psi_{j_0k}^{(2)}(x,y) + \sum_{k\in\mathbb{Z}^2}\beta_{j_0k}^{(3)}\psi_{j_0k}^{(3)}(x,y)\right).$$

For more details about $D_{j_0}f(x,y)$ and the functions ψ_{j_0k}, see [14]. The coefficients α_{j_0k} and $\beta_{j_0k}^{(1)}$, $\beta_{j_0k}^{(2)}$, $\beta_{j_0k}^{(3)}$ are unique for every $j_0 \in \mathbb{N}$. The function ϕ_{j_0k} is defined as $\phi_{j_0k_1k_2}(x,y) = \phi_{j_0k_1}(x)\phi_{j_0k_2}(y)$.

Some important cases of wavelets are the Haar, Daubechies, Shannon, Meyer, and Morlet wavelets (see [8]). We used the Haar and Daubechies wavelets in our simulation studies. Accordingly, by equation (1) the copula density c can be expanded with

$$\alpha_{j_0k} = \int_{(0,1)^2} c(u,v)\phi_{j_0k}(u,v)\,du\,dv, \quad k \in \mathbb{Z}^2.$$

By the change of variables $u = F_1(t_1)$ and $v = F_2(t_2)$ we get

$$\alpha_{j_0k} = \int \phi_{j_0k}\big(F_1(t_1), F_2(t_2)\big)f(t_1,t_2)\,dt_1\,dt_2 = E_f\big(\phi_{j_0k}\big(F_1(T_1), F_2(T_2)\big)\big). \tag{2}$$

Assuming that $I = I(Y_{1i} \leq T, T_{2i} \leq T)$, when F_1 and F_2 (the marginal distributions) are known, a moment-based estimator of α_{j_0k} based on censored data is then given by

$$\begin{aligned}\hat{\alpha}_{j_0k} &= \int I\phi_{j_0k}\big(F_1(y_1), F_2(t_2)\big)\hat{F}(dy_1, dt_2)\\ &= \frac{1}{n}\sum_{i=1}^{n} W_{in}^{(\cdot)} I(Y_{1i} \leq t_1, T_{2i} \leq t_2)\phi_{j_0k}\big(F_1(Y_{1i}), F_2(T_{2i})\big).\end{aligned} \tag{3}$$

Then the wavelet-based estimator of c is given by

$$\hat{c_{j_0}}(u,v) = \sum_{k\in\mathbb{Z}^2}\hat{\alpha}_{j_0k}\phi_{j_0k}(u,v), \quad u,v \in (0,1).$$

When F_1 and F_2 are unknown, their empirical distribution functions F_{1n} and F_{2n} are used. So the rank-based estimator is as follows:

$$\begin{aligned}\tilde{\alpha}_{j_0k} &= \int I\phi_{j_0k}\big(F_{1n}(y_1), F_{2n}(t_2)\big)\hat{F}(dy_1, dt_2)\\ &= \frac{1}{n}\sum_{i=1}^{n} W_{in}^{(\cdot)} I(Y_{1i} \le t_1, T_{2i} \le t_2)\phi_{j_0k}\big(F_{1n}(Y_{1i}), F_{2n}(T_{2i})\big).\end{aligned} \tag{4}$$

Now we can introduce the linear wavelet-based estimator of c based on the ranks as

$$\tilde{c}_{j_0}(u,v) = \sum_{k\in\mathbb{Z}^2} \tilde{\alpha}_{j_0k}\phi_{j_0k}(u,v), \quad u,v \in (0,1). \tag{5}$$

When we have no censoring, the definition reduces to that of the copula estimator introduced in [9]. We further denote by K any constant that may change from one line to another, does not depend on j, k, n, but depends on the wavelet basis and on $\|c\|_\infty = \sup_{(u,v)\in(0,1)} |c(u,v)|$ and $\|c\|_2 = \int c(u,v)^2\,du\,dv$.

Since we deal with the wavelet method, it is very common to consider Besov spaces as functional spaces because they are characterized in terms of wavelet coefficients as follows. Besov spaces depend on three parameters $s > 0$, $1 < p < \infty$, and $1 < q < \infty$ and are denoted by B^s_{pq}. Let $f \in L_2(\mathbb{R}^d)$ (that is, $L_2(\mathbb{R}^2)$ in our paper), and let $s < r$ (wavelet regularity). Define the sequence norm of the wavelet coefficients of a function $f \in B^s_{pq}$ by

$$|f_{B^s_{pq}}| = \left(\sum_{k\in\mathbb{Z}^2} |\alpha_{j_0k}|^p\right)^{1/p} + \left(\sum_{j\ge j_0}\left[2^{j_0(s+d(1/2-1/p))}\left(\sum_{k\in\mathbb{Z}^2} |\beta_{j,k}|^p\right)^{1/p}\right]^q\right)^{1/q},$$

where $(|\beta_{j,k}|^p)^{1/p} = (\sum_{k\in\mathbb{Z}^2}\sum_{\epsilon\in S_2} |\beta^{\epsilon}_{j,k}|^p)^{1/p}$. We assume that the copula function c belongs to a Besov space.

3 Main results

In this section, we present the main results. The following theorems show that the wavelet-based estimators based on censoring attain nearly optimal convergence rates over a large range of Besov function classes. We also show that our estimator obtains the optimal convergence rates under mean integrated squared error (MISE) by accepting some mild conditions. The mean integrated squared error (MISE) is defined by

$$\operatorname{MISE}(\tilde{c}_{j_0}, c) = E_f\left[\int_0^1\int_0^1 \left\{\tilde{c}_{j_0}(u,v) - c(u,v)\right\}^2 dv\,du\right].$$

In view of decomposition (1) for c, it is obvious that

$$\operatorname{MISE}(\tilde{c}_{j_0}, c) = \operatorname{MISE}(\tilde{c}_{j_0}, c_{j_0}) + \int_0^1\int_0^1 \left\{D_{j_0}c(u,v)\right\}^2 dv\,du.$$

The bias term in the equation can be bounded as in the proof of Lemma 1 in [4],

$$\int_0^1\int_0^1 \left\{D_{j_0}c(u,v)\right\}^2 dv\,du \le M2^{-2j_0s^*}.$$

Precisely, suppose that c belongs to the ball of radius $M > 0$ in the Besov space $B^s_{p,q}(M)$, where the parameters $s > 0$ and $p \geq 2$, $1 \leq q \leq \infty$ or $s > 2/p - 1$ and $p \in [1,2]$, $1 \leq q \leq \infty$; also, s^* can be defined as $s^* = s + 1 - 2/p$ if $p \in [1,2]$ and $s^* = s$ otherwise. Also, by defining $\hat{c}_{j_0}$ based on Eq. (3), we can easily show that

$$\operatorname{MISE}(\tilde{c}_{j_0}, c_{j_0}) \leq 2\operatorname{MISE}(\tilde{c}_{j_0}, \hat{c}_{j_0}) + 2\operatorname{MISE}(\hat{c}_{j_0}, c_{j_0}).$$

We now look for an optimal upper bound for each of the above sequences.

Lemma 1 *Let $\hat{\alpha}$ be as in* (3), *and let j_0 be an arbitrary integer in $\mathbb{N}$. Then*

$$S_1 \equiv E\sum_{k\in z^2} (\hat{\alpha}_{j_0k} - \alpha_{j_0k})^2 \leq K_1 \frac{2^{2j_0}}{n}$$

for some constant $K_1 > 0$ depending only on ϕ and either

$$\|c\|_2 = \int c(u,v)^2 \, du \, dv \quad \textit{or} \quad \|c\|_\infty = \sup_{u,v\in(0,1)} \bigl|c(u,v)\bigr|.$$

Also note that the proposed linear wavelet estimator of c_{j_0} is $\hat{c}_{j_0}(x,y) = \sum_{k\in\mathbb{Z}^2} \hat{\alpha}_{j_0k} \times \phi_{j_0k}(x,y)$. Then

$$\operatorname{MISE}(\hat{c}_{j_0}, c_{j_0}) = O\biggl(\frac{2^{2j_0}}{n}\biggr).$$

Note that j_0 must be chosen so that $2^{j_0} \ll \sqrt{n}$.

Proof of Lemma 1 The proof is similar to optimality results in Sect. 5 of [14]. □

Theorem 1 *Assume that the function ϕ is m-differentiable, and let $\tilde{c}_{j_0}$ be the copula density estimator of c defined in* (5). *Then there exists a constant $K_1 > 0$ such that, for a given $(u,v) \in (0,1)^2$ and any level j_0 satisfying $2^{j_0} \simeq (n/\log n)^{1/2-1/2m}$,*

$$\operatorname{MISE}(\tilde{c}_{j_0}, \hat{c}_{j_0}) \leq K_1 \frac{2^{2j_0}}{n}\biggl(2^{2j_0}\frac{\log n}{n} + 2^{-j_0}\log n\biggr).$$

The error term associated with the use of ranks is negligible with respect to the usual error term as soon as $2^{j_0} \gg \log(n)$. For the proof, see the Appendix.

Now, combining the results of Lemma 1 and Theorem 1, we are in a position to express the main theorem of this section.

Theorem 2 *Let ϕ be a scaling function mentioned in Sect.* 2.2 *having m derivatives with $m > 1 + 1/s^*$, and for arbitrary $j_0 \in \mathbb{N}$, let $\tilde{c}_{j_0}$ be the estimator in* (5). *Then there exists a constant $K_2 > 0$ such that, for all $M \in (0,\infty)$, $s > 2/p - 1$, and $p, q \in [1,\infty)$, if j_0 satisfies $2^{j_0} \simeq n^{1/2+2s^*}$, then*

$$\sup_{c\in B^s_{p,q}(M)} \operatorname{MISE}(\tilde{c}_{j_0}, c) \leq K_2 n^{-s^*/1+s^*}.$$

Note that the procedure of estimation described is optimal on the Besov space. The simulation studies are applied in the next section. For the proof, see the Appendix.

Remark 1 This study can be regarded as an extension from complete data to randomly right censored data. If we assume that there is no censoring, that is, $G \equiv 0$ on $(-\infty, +\infty)$, then $\delta_{1i} = 1$ for all $i = 1, 2, \ldots, n$, and the estimators defined by (4) and (3) are the same as those of [14] for the complete data case. Therefore our estimators can be regarded as an extension of those of [14] from complete data to randomly right censored data.

4 Simulation studies and analysis for real data

In this section, we conduct simulation studies to investigate the performance of the proposed estimator (5) for censored data and compare it with the kernel estimator by an average squared error. Wavelet estimators for the copula, for noncensored data, have been proposed in [14].

4.1 Simulation study

The simulation scheme consisted in three steps.

Step 1: We simulate a random sample of size $n = 2000$ to display some scatter and rank–rank plots according to the following scheme.

(i) The marginal distribution of random variables T_1 and T_2 (F_1 and F_2) are simulated from three distributions: (a) F_1 is the B distribution with shape parameter $\alpha_1 = 0.5$ and scale parameter $\alpha_2 = 0.7$, F_2 is the exponential distribution with mean $1/3$; (b) F_1 is $N(0,1)$, and F_2 is as the same as in part (a); and (c) both F_1 and F_2 are uniform distributions.

(ii) For the dependence structure, we consider two copula families, Gumbel and Clayton with Kendall's tau $\tau = 0.5$.

(iii) Censoring variables were simulated from $\mathrm{Exp}(\nu(x))$ where $\nu(x) = x + 0.7$ with results in approximately 0.40 censoring.

The plots are shown in Figs. 1 and 2. In both figures the scatter plots in the top row show what happens when F_2 is exponential with mean $1/3$ and F_1 is either $B(0.5, 0.7)$ or $N(0,1)$. In the bottom row the left one shows the same scatter plot, but the margins are considered as $U(0,1)$, and the right one shows the pairs of normalized ranks. It is worth mentioning that Fig. 1 displays scatter plots generated from the Gumbel copula, and Fig. 2 displays the same ones from the Clayton copula. As the number of observations is increased, the rank–rank plots are more unreadable. It is clear that for a random sample of size $n = 2000$ or more, the square in these plots could be completed filled, and all features of distribution had been lost. As a suggestion, we would consider the plots of the empirical copula function of the pairs $(R_i/n, S_i/n)$. For the same data as in panel (d, bottom right) in Figs. 1 and 2, 3d-histograms of the relative frequencies of the pseudo-observations $(R_i/n, S_i/n)$ are illustrated in Fig. 3. The plots show the relative frequency of $n = 2000$ pairs $(R_i/n, S_i/n)$ in a 32×32 regular partitions of the unit square for Gumbel (left column) and Clayton (right column) copulas in two parts, full data (top row) and censored data (bottom row) with Kendall's tau $\tau = 1/2$.

Step 2: Marginal distribution of T_1 and T_2 are simulated from $B(0.5, 0.7)$ and $\mathrm{Exp}(3)$, respectively. Then for $N = 2^J \leq \sqrt(n) < 2^{J+1}$, the empirical scaling coefficients at resolution

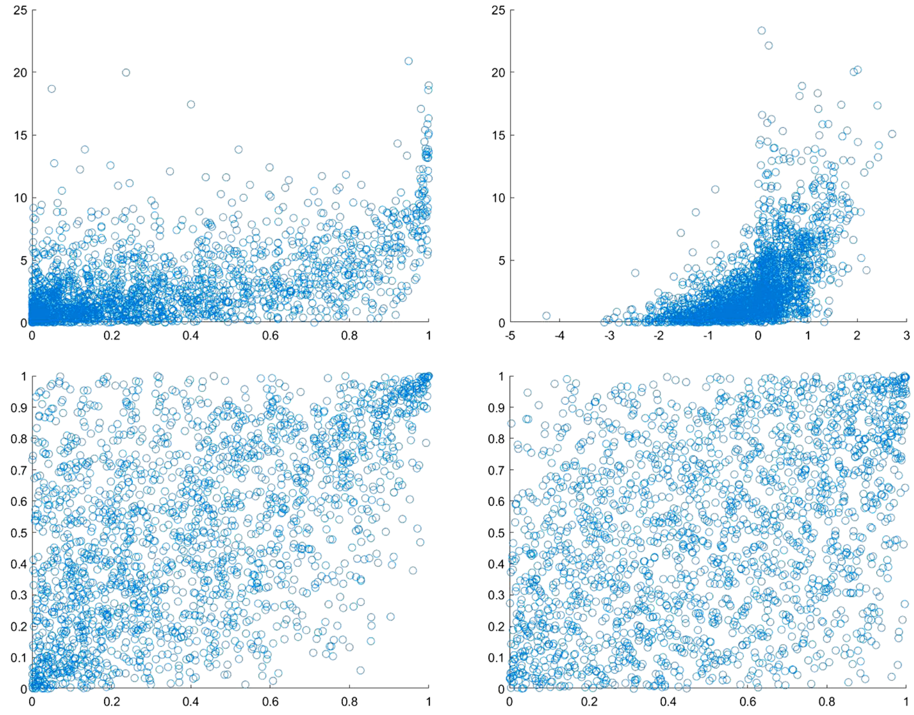

Figure 1 Scatter and rank-rank plots for a sample of size 2000 from the Gumbel copula based on censored data with $\tau = 1/2$, once the margins have been transformed to F_1 and F_2. (**a**, top left): $F_1 = B(0.5, 0.7)$ and $F_2 = \varepsilon(3)$; (**b**, top right): $F_1 = N(0,1)$ and $F_2 = \varepsilon(3)$; (**c**, bottom left): $F_1 = F_2 = U(0,1)$; (**d**, bottom right): the rank-rank plot for Gumbel copula

level J are computed as follows:

$$\tilde{\alpha}_{Jk_1k_2} = \frac{1}{n}\sum_{i=1}^{n} w_{in} I\left\{\frac{k_1-1}{N} < \frac{R_i}{n} \leq \frac{k_1}{N}, \frac{k_2-1}{N} < \frac{S_i}{n} \leq \frac{k_2}{N}\right\}$$

for all $k_1, k_2 \in \{1, \ldots, N\}$. Finally, we obtain the wavelet estimators for Gumbel and Clayton copulas (equation (5)) and consider their plots at levels $J-1$ and $J-2$. The graphs are in Fig. 4.

Step 3: We find the average squared error (ASE) calculated for several wavelet copula estimators, including (5) for different sample sizes n. The results for ASE and the total number of replications $N = 100$ for two different dependence copula parameters and for wavelet and kernel methods are shown in Table 1. The ASE criterion is defined as

$$ASE = \frac{1}{N}\sum_{l=1}^{N}\left(\frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}\left(\tilde{c}^{(l)}(u_i, v_j) - c(u_i, v_j)\right)^2\right),$$

where $\tilde{c}^{(l)}$ denotes an estimator of c at the lth replication. In this simulation study, we have used Daubechies's compactly supported wavelet symmlet8 and level $j_0 = 5$.

The codes are in MATLAB environment using the Wavelab software. Our simulation study is done for sampling in three sizes $n = 100$, $n = 300$, and $n = 500$. The main result in the table is consistent with the conclusions of [14]. In both of these examples the sim-

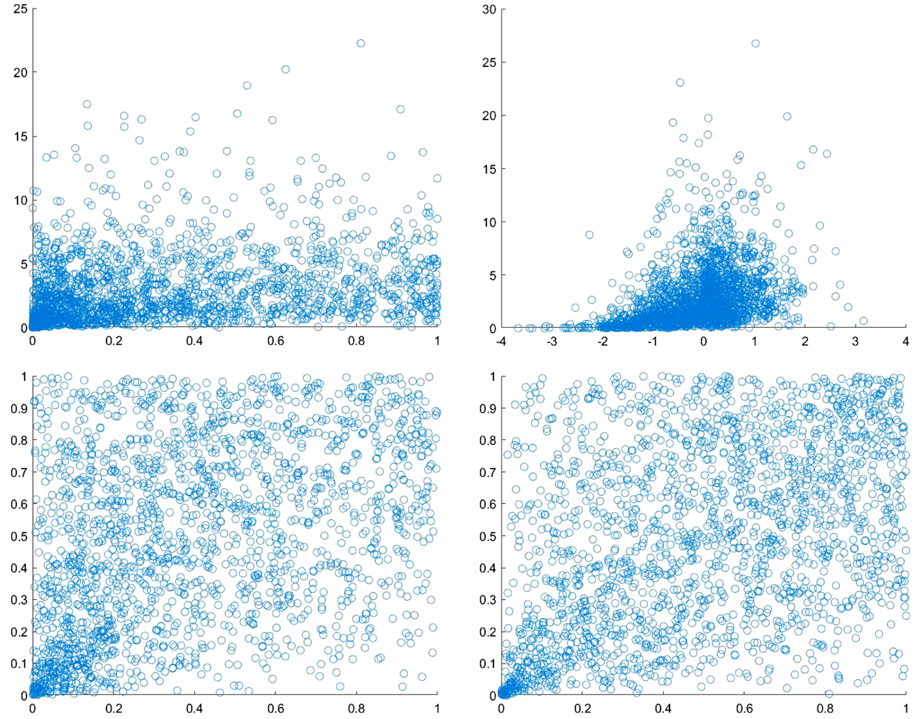

Figure 2 Scatter and rank-rank plots for a sample of size 2000 from the Clayton copula based on censored data with $\tau = 1/2$, once the margins have been transformed to F_1 and F_2. (**a**, top left): $F_1 = B(0.5, 0.7)$ and $F_2 = \varepsilon(3)$; (**b**, top right): $F_1 = N(0, 1)$ and $F_2 = \varepsilon(3)$; (**c**, bottom left): $F_1 = F_2 = U(0, 1)$; (**d**, bottom right): the rank-rank plot for Clayton copula

ulation results show that the wavelet estimator performs better than kernel estimators in terms of AMSE criterion. We can do the same work for some other copula functions like Gaussian, Frank, and Student copulas or for other wavelets such as Haar or Adelson wavelets.

4.2 Real data

Here we present an application of the proposed methodology for the data of [13], where the first observation is censored. The data consist of the indemnity payment (LOSS) and the allocated loss adjustment expense (ALAE) for 1500 general liability claims. The graphical representation of this data is considered in Fig. 5. Panels (a) and (b) show logarithm scale of the original data and the rank–rank plot, respectively. In panel (c) the 3D-histogram of Loss (censored) and Alae data is shown based on 16 × 16 grid. Panel (d) shows the wavelet-based estimator described in Sect. 2.

Due to our simulation results, the nonparametric procedure tries to provide a smooth copula estimate, and the wavelet-based estimator is a good estimator under the fact that a considerable copula family is Gumbel. Many authors who used this data in their researches claimed that the best representation of Loss and Alae data is the Gumbel copula.

5 Conclusions

Here we consider wavelet-based identification and estimation of a censored copula density function under a rank-based producer. We proposed a linear wavelet estimator and provided its asymptotic formulae for mean integrated square error. The wavelet methods

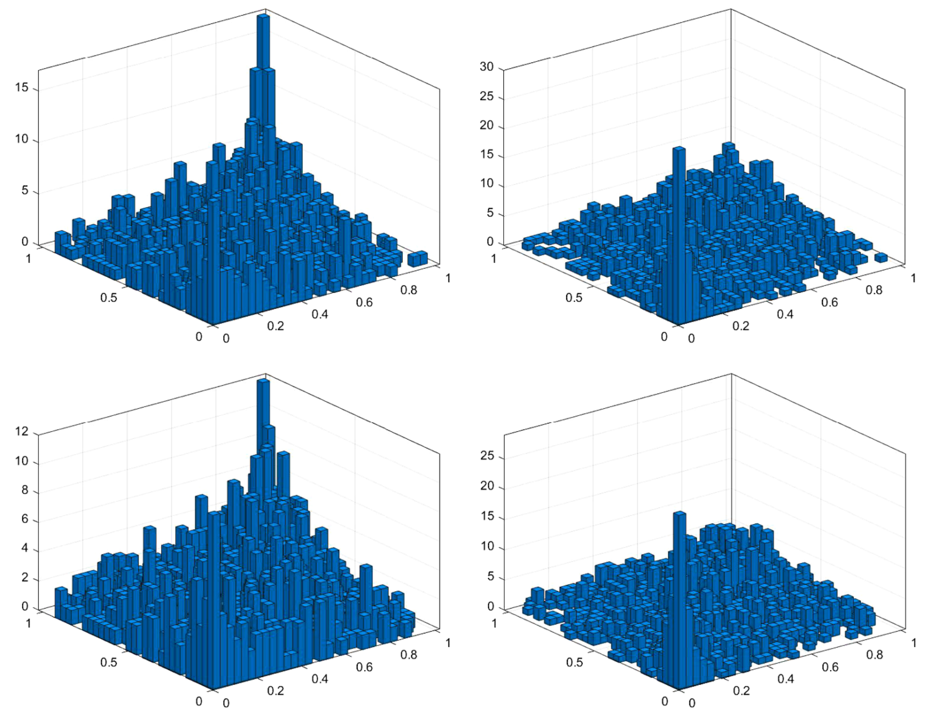

Figure 3 3d-histograms. (**a**, top left): Gumbel (full data), (**b**, top right): Clayton (full data), (**c**, bottom left): Gumbel (censored data), (**d**, bottom right): Clayton (censored data)

offer fast computations and easy updating in addition to being easily adapted to the design. We derived an analog of the asymptotic formula of the mean integrated square error in the context of kernel density estimators for censored data, admitting an expansion with distinct squared bias and variance components.

The numerical performance of the proposed linear wavelet density estimators was illustrated on simulated datasets. Comparisons between full data and censored data for some different sample sizes were also given. Although using wavelet-based estimator of a copula function is very useful for underlying dependence structure, it does not cover the main conditions of parametric models. In future work we might also consider using a nonlinear wavelet-based copula density estimator for randomly censored data or using for goodness-of-fit testing.

Appendix: Proofs

A.1 Proof of Theorem 1

The proofs follow along with the lines of that in [14] for the density function. Compared to the uncensored case, the main difficulty here is to handle the weights W_{in}. For this purpose, we use some assumptions of [15]. This is the direct approach to the copula function estimation problem under censoring. To complete the proof, we need some preparations. Suppose that Assumption 4 in [15] is satisfied. Here this assumption for our results is as follows.

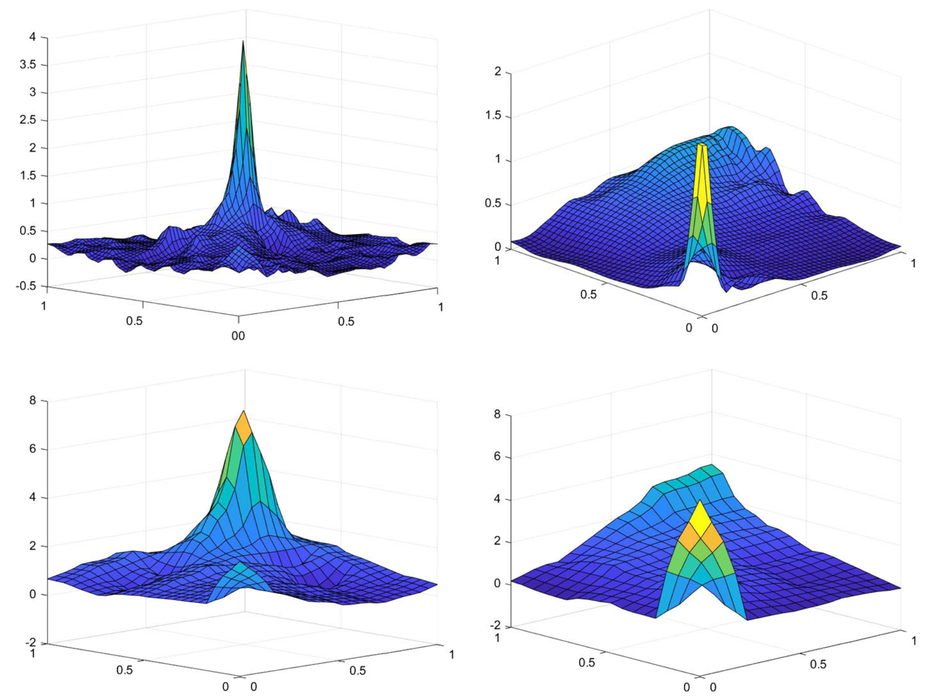

Figure 4 Wavelet smoothing of $n = 2000$ pairs of normalized ranks from Gumbel (left column) and Clayton (right column). (**a**, top left) and (**b**, top right) show wavelet estimators under censored data at level $j = J - 1$ with $J = 5$ for Gumbel and Clayton copula, respectively. (**c**, bottom left) and (**d**, bottom right) show wavelet estimators under censored data at level $j = J - 2$ with $J = 5$ for Gumbel and Clayton copulas, respectively

Table 1 Computed value for ASE compared between the proposed wavelet estimator in our paper and kernel estimator (defined in lopez(2015)) based on two copulas for various sample sizes

Estimation methods	ASE		
	$n = 100$	$n = 300$	$n = 500$
Wavelet Gumbel ($\tau = 0.25$)	1.07	1.25	1.36
Kernel ($h = 0.4$) Gumbel ($\tau = 0.25$)	1.32	1.45	1.57
Wavelet Clayton ($\tau = 0.25$)	1.15	1.82	2.02
Kernel ($h = 0.4$) Clayton ($\tau = 0.25$)	1.37	2.01	2.32
Wavelet Gumbel ($\tau = 0.75$)	6.13	7.32	7.59
Kernel ($h = 0.4$) Gumbel ($\tau = 0.75$)	6.63	7.66	7.95
Wavelet Clayton ($\tau = 0.75$)	7.20	8.74	9.11
Kernel ($h = 0.4$) Clayton ($\tau = 0.75$)	7.90	9.31	10.32

Assumption 1 Assume that $E(\frac{\delta_1}{1-G})^2 = E(\delta_1 g(T_1))^2 < \infty$, and assume that there exist i.i.d. random variables (Z_i) such that $\sup|W_{in} - W_i| \leq B_n Z_i$, where $B_n = \sup|\frac{\hat{G}-G}{1-\hat{G}}| = O(n^{-1/2})$ and $E[Z_i] = E[\frac{\delta_1}{1-G(Y_{1i})}] < \infty$.

Also, suppose that the function ϕ defined in Sect. 2.2 is m-differentiable. We define

$$\xi_k(Y_{1i}, T_{2i}) = \bigl(\phi_{j_0k}\bigl(F_{1n}(Y_{1i}), F_{2n}(T_{2i})\bigr) - \phi_{j_0k}\bigl(F_1(Y_{1i}), F_2(T_{2i})\bigr)\bigr)I_{(Y_{1i} \leq A_1, Y_{2i} \leq A_2)}.$$

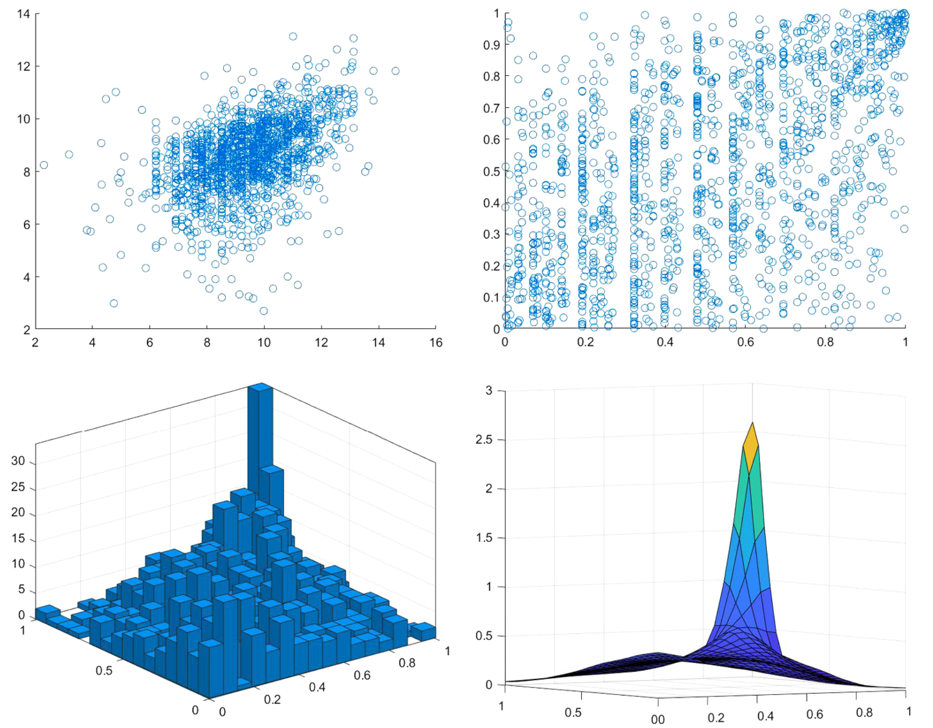

Figure 5 (**a**, top left): original data on the logarithm scale, (**b**, top right): rank–rank plot, (**c**, bottom left): 3D-histogram, panel (**d**, bottom right): wavelet-based copula density estimate

In addition, $\tilde{\alpha}_{j_0k} - \hat{\alpha}_{j_0k} = \frac{1}{n}\sum_{i=1}^{n} W_{in}\xi_k(Y_{1i}, T_{2i})$. So we obtain

$$
\begin{aligned}
E\biggl(\sum_k (\tilde{\alpha}_{j_0k} - \hat{\alpha}_{j_0k})^2\biggr) &= \sum_k E\Biggl(\frac{1}{n}\sum_{i=1}^{n} W_{in}\xi_k(Y_{1i}, T_{2i})\Biggr)^2 \\
&= \sum_k E\Biggl(\frac{1}{n}\sum_{i=1}^{n} W_i\xi_k(Y_{1i}, T_{2i}) + \frac{1}{n}\sum_{i=1}^{n}(W_{in} - W_i)\xi_k(Y_{1i}, T_{2i})\Biggr)^2 \\
&\leq 2\Biggl(\sum_k E\Biggl(\frac{1}{n}\sum_{i=1}^{n} W_i\xi_k(Y_{1i}, T_{2i})\Biggr)^2 \\
&\quad + \sum_k E\Biggl(\frac{1}{n}\sum_{i=1}^{n}(W_{in} - W_i)\xi_k(Y_{1i}, T_{2i})\Biggr)^2\Biggr) \\
&=: 2(T_1 + T_2).
\end{aligned}
$$

First, following the proof of Proposition 1 in [14], we obtain a bound for T_1:

$$
\begin{aligned}
T_1 \leq{}& \frac{1}{n^2}\sum_k \sum_{i\in I(j,0)\cup I_1(j,\epsilon)} E\bigl(W_i\xi_k(Y_{1i}, T_{2i})\bigr)^2 \\
&+ \frac{1}{n^2}\sum_k \sum_{i\neq l\in I(j,0)\cup I_1(j,\epsilon)} \bigl|E\bigl(W_iW_l\xi_k(Y_{1i}, T_{2i})\xi_k(Y_{1l}, T_{2l})\bigr)\bigr| + K2^{2j}n^{-\delta}E(W_i)^2.
\end{aligned}
$$

Due to the last sentence in Sect. 2.1 and because of the limit function g, we get

$$E\big(W_i\xi_k(Y_{1i},T_{2i})\big)^2 = gE\big(\xi_k(T_{1i},T_{2i})\big)^2 \le K_1E\big(\xi_k(T_{1i},T_{2i})\big)^2$$

and also

$$\begin{aligned} E\big(W_iW_l\xi_k(Y_{1i},T_{2i})\xi_k(Y_{1l},T_{2l})\big) &= gE\big(\xi_k(T_{1i},T_{2i})\xi_k(T_{1l},T_{2l})\big) \\ &\le K_1E\big(\xi_k(T_{1i},T_{2i})\xi_k(T_{1l},T_{2l})\big). \end{aligned}$$

Similarly, we have $E(W_i)^2 = \frac{1}{1-G} = g \le K_1$.

Since the support of the scaling function is compact, we finally obtain the bound for T_1:

$$\begin{aligned} T_1 &\le KK_1\frac{1}{n^2}2^{2j}\big(n2^{-j}\big)2^{3j}\frac{\log(n)}{n} + KK_1\frac{1}{n^2}2^{2j}\big(n2^{-j}\big)^2 2^{3j}\frac{\log(n)}{n} + KK_1n^{-\delta+1} \\ &\le K\frac{2^{2j}}{n}\left(2^{2j}\frac{\log(n)}{n} + 2^{-j}\log(n)\right). \end{aligned}$$

In addition, according to Assumption 1, we have

$$\begin{aligned} E\big((W_{in}-W_i)\xi_k(Y_{1i},T_{2i})\big)^2 &= E\left(\frac{\delta_1(\hat{G}-G)}{(1-\hat{G})(1-G)}\xi_k(Y_{1i},T_{2i})\right)^2 \\ &= \frac{(\hat{G}-G)^2}{(1-\hat{G})^2}E\big(Z_i\xi_k(T_{1i},T_{2i})\big)^2 \\ &\le Kn^{-1}E\big(\xi_k(T_{1i},T_{2i})\big)^2. \end{aligned}$$

Similarly, the bound for $(W_{in}-W_i)\xi_k(Y_{1i},T_{2i})$ follows:

$$\begin{aligned} &E\big((W_{in}-W_i)(W_{ln}-W_l)\xi_k(Y_{1i},T_{2i})\xi_k(Y_{1l},T_{2l})\big) \\ &\quad= \frac{(\hat{G}-G)^2}{(1-\hat{G})^2}E\big(Z_i^2\xi_k(T_{1i},T_{2i})\xi_k(T_{1l},T_{2l})\big) \\ &\quad\le Kn^{-1}E\big(\xi_k(T_{1i},T_{2i})\xi_k(T_{1l},T_{2l})\big). \end{aligned}$$

It remains to find the bound for the last part of T_2:

$$E(W_{in}-W_i)^2 = E\left(\frac{\delta_1(\hat{G}-G)}{(1-\hat{G})(1-G)}\right)^2 = \frac{(\hat{G}-G)^2}{(1-\hat{G})^2}E(Z_i)^2 \le Kn^{-1},$$

where the last equality follows from Assumption 1. At the end, the sharper bound for C_2 is

$$\begin{aligned} T_2 \le{}& \frac{1}{n^2}\sum_k\sum_{i\in I(j,0)\cup I_1(j,\epsilon)}\frac{(\hat{G}-G)^2}{(1-\hat{G})^2}E\big(Z_i\xi_k(T_{1i},T_{2i})\big)^2 \\ &+ \frac{1}{n^2}\sum_k\sum_{i\neq l\in I(j,0)\cup I_1(j,\epsilon)}\frac{(\hat{G}-G)^2}{(1-\hat{G})^2}\big|E\big(Z_i^2\big)\xi_k(T_{1i},T_{2i})\xi_k(T_{1l},T_{2l})\big| \end{aligned}$$

$$
\begin{aligned}
&+ Kn^{-\delta} 2^{2j} \frac{(\hat{G} - G)^2}{(1 - \hat{G})^2} E(Z_i)^2 \\
&\leq KK_1 n^{-1} \frac{1}{n^2} 2^{2j} \bigl(n2^{-j}\bigr) 2^{3j} \frac{\log(n)}{n} + KK_1 n^{-1} \frac{1}{n^2} 2^{2j} \bigl(n2^{-j}\bigr)^2 2^{3j} \frac{\log(n)}{n} + KK_1 n^{-1} n^{-\delta+1} \\
&\leq K \frac{2^{2j}}{n^2} \biggl(2^{2j} \frac{\log(n)}{n} + 2^{-j} \log(n) \biggr).
\end{aligned}
$$

Therefore with the bounds for T_1 and T_2 the proof of Theorem 1 is complete.

A.2 Proof of the Theorem 2

Combining the result of Lemma 1 and Theorem 1 and the fact that the error term associated with the use of ranks in the proof of Theorem 1 is negligible with respect to the usual error term as soon as $2^{j_0} \gg \log(n)$, the proof is complete.

Funding
Not applicable.

Availability of data and materials
The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
This work was carried out in collaboration among all authors. All authors read and approved the final manuscript.

Author details
[1]Department of Statistics, Payame Noor University, Tehran, Iran. [2]Department of Statistics, Payame Noor University, Mashhad, Iran. [3]Faculty of Science, Gonbad Kavous University, Gonbad Kavous, Iran.

Publisher's Note

Received: 6 March 2019 Accepted: 21 June 2019 Published online: 05 July 2019

References

1. Autin, F., Lepennec, E., Tribouley, K.: Thresholding methods to estimate the copula density. J. Multivar. Anal. **101**, 200–222 (2010). http://doi.org/10.1016/j.jmva.2009.07.009
2. Bouye, E., Durrleman, V., Nikeghbali, A., Riboulet, G., Roncalli, T.: Copulas for finance: a reading guide and some applications. In: Groupe de Recherche Operationelle. Credit Lyonnais, Paris (2000)
3. Bouye, E., Durrleman, V., Nikeghbali, A., Riboulet, G., Roncalli, T.: Copulas for finance—A reading guide and some applications. Social Science Research Network Working Paper Series (2007). https://doi.org/10.2139/ssrn.1032533
4. Butucea, C., Tribouley, K.: Nonparametric homogeneity tests. J. Stat. Plan. Inference **136**, 597–639 (2006). http://doi.org/10.1016/j.jspi.2004.08.003
5. Chatrabgoun, O., Parham, G.: Copula density estimation using multiwavelets based on the multiresolution analysis. Commun. Stat., Simul. Comput. **45**, 3350–3372 (2014)
6. Chatrabgoun, O., Parham, G., Chinipardaz, R.: A Legendre multiwavelets approach to copula density estimation. Stat. Pap. **58**, 673–690 (2017). http://doi.org/10.1007/s00362-015-0720-0
7. Cherubini, U., Luciano, E., Vecchiato, W.: Copula Methods in Finance. Wiley Finance Series. Wiley, Chichester (2004)
8. Daubechies, I.: Ten Lectures on Wavelets. SIAM, Philadelphia (1992)
9. Deheuvels, P.: La fonction de dépendance empirique et ses propriétés: Un test non paramétrique d'indépendance. Bull. Cl. Sci., Acad. R. Belg. (5) **65**, 274–292 (1979)
10. Embrechts, P., Kluppelberg, C., Mikosch, T.: Modeling Extremal Events for Insurance and Finance. Springer, Berlin (1997)
11. Fermanian, J.-D., Wegkamp, M.: Time dependent copulas. Preprint (2004). http://doi.org/10.1.1.332.6192
12. Fleming, T.R., Harrington, D.P.: Counting Processes and Survival Analysis. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. Wiley, New York (1991). http://doi.org/10.1002/9781118150672
13. Frees, E., Valdez, E.A.: Understanding relationships using copulas. N. Am. Actuar. J. **2**, 1–25 (1998). http://doi.org/10.1080/10920277.1998.10595667
14. Genest, C., Masiello, E., Tribouley, K.: Estimating copula densities through wavelets. Insur. Math. Econ. **44**, 170–181 (2009). http://doi.org/10.1016/j.insmatheco.2008.07.006

15. Gribkova, S., Lopez, O.: Non-parametric copula estimation under bivariate censoring. Scand. J. Stat. **42**, 925–946 (2015). http://doi.org/10.1111/sjos.12144
16. Joe, H.: Multivariate Models and Dependence Concepts. Chapman & Hall, London (1997). https://doi.org/10.1002/(SICI)1097-0258(19980930)17:18<2154::AID-SIM913>3.0.CO;2-R
17. Li, L.: Non-linear wavelet-based density estimators under random censorship. J. Stat. Plan. Inference **117**, 35–58 (2003). http://doi.org/10.1016/so378-3758(02)00366-x
18. Meyer, Y.: Wavelets: Algorithms and Applications. SIAM, Philadelphia (1993). http://doi.org/10.1137/1036136
19. Morettin, P.A., Toloi, C.M.C., Chiann, C., de Miranda, J.C.S.: Wavelet smoothed empirical copula estimators. Braz. Rev. Finance **8**, 263–281 (2010)
20. Nelsen, R.B.: An Introduction to Copulas, 2nd edn. Springer, New York (2006)
21. Patton, A.J.: A review of copula models for economic time series. J. Multivar. Anal. **110**, 4–18 (2012). http://doi.org/10.1016/j.jmva.2012.02.021
22. Sklar, A.: Fonctions de répartition à *n* dimensions et leurs marges. Publ. Inst. Stat. Univ. Paris **8**, 229–231 (1959)
23. Van der Laan, M.J.: Efficient estimation in the bivariate censoring model and repairing NPMLE. Ann. Stat. **24**(2), 596–627 (1996). http://jstor.org/stable/2242663
24. Wang, W., Wells, M.T.: Nonparametric estimators of the bivariate survival function under simplified censoring conditions. Biometrika **84**(4), 863–880 (1997). http://doi.org/10.1093/biomet/84.4.863